Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Hybrid Feature Vector Extraction in Unsupervised Learning Neural Classifier

P. S. Kostka, E. J. Tkacz, *IEEE Member*, D. Komorowski

Abstract – **Feature extraction and selection method as a preliminary stage of heart rate variability (HRV) signals unsupervised learning neural classifier is presented. Multi-domain, mixed new feature vector is created from time, frequency and time-frequency parameters of HRV analysis. The optimal feature set for given classification task was chosen as a result of feature ranking, obtained after computing the class separability measure for every independent feature. Such prepared a new signal representation in reduced feature space is the input to neural classifier based on introduced by Grosberg Adaptive Resonance Theory (ART2) structure. Test of proposed method carried out on the base of 62 patients with coronary artery disease divided into learning and verifying set allowed to chose these features, which gave the best results. Classifier performance measures obtained for unsupervised learning ART2 neural network was comparable with these reached for multiplayer perceptron structures.**

I. Introduction

Feature extraction consisting in revealing the most significant information is the preliminary process, in many classification problems belonging to the group of computer-aided biomedical applications. Originally recorded raw data from patient create rather high dimensional *N*-element feature vector $x_I \in X \subseteq \Re^N$, able to describe whole complex objects or analysed system. For a given narrow classification task it almost always consists of redundant components, which can make the whole input signal clustering procedure more difficult. The role of feature extraction and selection stages is just to expose the most discriminative elements from the input feature set and discard remain, reducing also the classifier complexity.

Regarding the application field of described classifier a new feature vector created as a combination of time (T), frequency(F) and mixed time-frequency (T-F) parameters obtained from original HRV signal was proposed. Assuming, that important HRV based features are characterized by local information in the duals domains of time and frequency and treating HRV as non-stationary signal from its nature, wavelet transform was chosen as T-F signal representation tool [1].

In next feature selection stage the most representative feature set from three groups (T, F and T-F features) is created based on feature ranking competition algorithm. In presented HRV classifier structure, dimension reduction method based on selecting the best feature subset according to assumed criteria were used. Because the evaluation of optimal cost function using probability of misclassification is too complex [2], in presented approach simpler criterion based on class separability (CS) were applied. The measure of "discriminant power" of each single feature was computed both based on Fisher's class separability index and entropy parameters [3]. From selected features, the most representative new vector was created, which is the input of neural classifier part of whole system.

On the last classifier stage of proposed system, after our trials with supervised learnt classifier structures based on multiplayer perceptron and radial basic functions [4] in this paper the application of clustering methods based on unsupervised learnt Adaptive Resonance Theory (ART) introduced by Carpenter and Grossberg is presented. So-called *data clustering* problem arises in a great variety of fields, including artificial intelligence, machine learning, and statistics. Assuming that the prepared features form a set of instances with an underlying group-structure, data clustering may be roughly defined as the search for the best description of this group-structure, when the true group membership of every instance is unobserved. In presented in this paper case, these structures fulfil the role of non-linear classifier of extracted HRV signal representation parameters in proposed method of screening examinations of coronary artery disease.

II. Methods

A. General structure of proposed method

Proposed system using time-frequency methods in feature extraction part and unsupervised learning neural networks as classifier of heart rate variability (HRV) signals consists of following stages:

I. HRV signal preprocessing:
- Continuous representation of HRV (Derivative Cubic Spline Interpolation method)
- HRV uniformly resampling (f_S = 5 [Hz]);

II. New feature vector extraction, using hybrid, multi-type features from three groups:

Manuscript received May 2, 2005. This work was supported in part by the State Committee for Scientific Research in Poland under Grant T11E 01925

P. S. Kostka. [1]Institute of Electronics, Division of Microelectronics and Biotechnology, Silesian University of Technology, Gliwice, Poland, pkostka@polsl.pl, [2]Medical University of Silesia, Faculty of Pharmacy and Laboratory Medicine,Department of Bionics, Sosnowiec, Poland.

E. J. Tkacz [1]Institute of Electronics, Division of Microelectronics and Biotechnology, Silesian University of Technology, Gliwice, Poland etkacz@polsl.pl, [2]Medical University of Silesia, Faculty of Pharmacy and Laboratory Medicine,Department of Bionics, Sosnowiec, Poland etkacz@slam.katowice.pl .

0-7803-8740-6/05/$20.00 ©2005 IEEE.

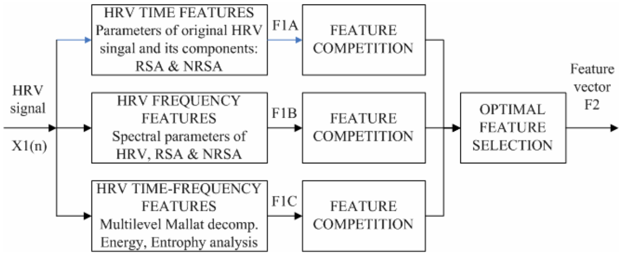

Fig.1 Structure of hybrid multi-type HRV feature extraction and selection algorithm.

- Time domain features: statistical parameters of original HRV signal and its breath related component Respiratory Sinus Arrhythmia (RSA) and NRSA reflecting influence of remain factors (autonomous nervous system) on HR modulation
- Frequency domain features: spectral parameters computed for analyzed HRV signal.
- Time-frequency features: T-F HRV representation, based on wavelet transform, which is suitable for non-stationary signals

III. Feature Selection, based on class separability index computed for every extracted feature.

IV. Classification – Supervisory learnt, nonlinear multilayer perceptron.

Decision rules, which assign the neural network outputs to pathological or physiological groups based on elaborated norms.

B. Feature extraction

A generalized feature extraction method can be expressed as a map f: $X_1 \rightarrow X_{F1}$, such that $X_{F1} \in \Re^M$ is the M-dimensional feature space , where M<<N.

As presented in fig.1 three feature sets based on HRV signal were computed.

I. The RSA and NRSA components were obtained from HRV and then set of statistical parameters (mean, standard variation and range) characterizing these signals were calculated to create the $F1_A$ feature vector:

$\boldsymbol{F1_A}$ = [HRV_M , HRV_{STD} , HRV_{RNG} , RSA_M , RSA_{STD} , RSA_{RNG} , $NRSA_M$, $NRSA_{STD}$, $NRSA_{RNG}$]

II. The energy of low (LF) and high frequency (HF) components of HRV spectrum as well as their ratio were included in $F1_B$ feature vector:

$\boldsymbol{F1_B}$ = [HRV_{LF_EN}, HRV_{HF_EN}, $HRV_{LF\_HF\_R}$]

III. The most complex feature set $\boldsymbol{F1}_C$ was created based on time-frequency (T-F) HRV analysis.

From several T-F signal analysis methods wavelet transform was chosen as a mathematical tool able to deal with non-stationarity HRV signal. Taking into consideration specific features of the HRV signal, especially that its significant frequency components are included in the range: $f_{HRV} \in <0;0.5>$ [Hz] , the grid of discrete wavelet scale – a values was created, corresponding to Mallat signal decomposition levels (for sampling frequency f_S = 5[Hz], six levels corresponding to scale values: a^i , i=3..8 were taken into consideration).

As a next step, to create the new features vector, for every signal component obtained on each decomposition level a set of parameters was computed. For each subspace wavelet coefficients were squared and normalised to obtain the energy probability distribution (1):

$$p_i = \frac{c_i^2}{\sum_{k=1}^{n} c_k^2} \qquad (1)$$

For each wavelet scale, the sorted series may be considered as an inverse empirical cumulative energy distribution function (ECDF). Base on this parameters the Shannon entropy E (2) of energy distribution p_i (1) were calculated as a measure of energy unpredictability in each wavelet decomposition subspace.

$$E = \sum_i p_i \log_2(p_i) \qquad (2)$$

This procedure allowed to reveal a group of new features based on energy and entropy measures.

The whole set of new feature vectors $\overline{F1_{C1}}$ .. $\overline{F1_{C5}}$, created as a result of multilevel Mallat signal decomposition, which is put to the input of classifier structure includes the following groups of parameters as a series for every of ith decomposition level:

I. Mean values of wavelet coefficients in each subband (frequency distribution information) - $\overline{F1_{C1}}$

II. Standard deviations of wavelet coefficients (level of change of frequency distribution information) - $\overline{F1_{C2}}$

III. Energy of ith component p_i (1) - $\overline{F1_{C3}}$

IV. Shannon entropy of wavelet component (distribution of the amount of information included in every subband) - $\overline{F1_{C4}}$

V. Shannon entropy E of energy distribution p_i (2) - $\overline{F1_{C5}}$

C. Feature selection

Feature set may be considered near to optimum if it minimizes chosen error based criterion function. There are two approaches to feature selection problem:

I. Feature subset selection

II. Feature projection, which tries to find optimal original feature combination (projection) into smaller set of new features. Principle component analysis (PCA) or projection pursuit [5] are often used feature projection methods.

In presented HRV classifier structure, dimension reduction method based on selecting the best feature subset according to assumed criteria were used. Because the evaluation of optimal cost function using probability of misclassification is too complex [2], in presented approach simpler criterion based on class separability (CS) were applied.

Considering a two class problem with an original M-dimensional feature set space X_{F1} a feature selection algorithm used in presented work is following:

I. Create two feature matrices M_{F1}^{p}, M_{F1}^{q}, representing two classes: *p* and *q* consisting of patterns (vectors) $x_{F1}^{(p,m)}$, $x_{F1}^{(q,m)}$ from learning data set:

$M_{F1}^{p} = \left[x_{F1}^{(p,1)}, x_{F1}^{(p,2)} \; \ldots \; x_{F1}^{(p,P)}\right]$ for class *p*, and $M_{F1}^{q} = \left[x_{F1}^{(q,1)}, x_{F1}^{(q,2)} \; \ldots \; x_{F1}^{(q,P)}\right]$ for class *q*, where: $x_{F1}^{(p,m)} = \left[x_{F1\_1}^{(p,m)}, x_{F1\_2}^{(p,m)} \; \ldots \; x_{F1\_M}^{(p,m)}\right]^T$ - m[th] pattern in class *p*.

II. Assuming, that we are trying to evaluate the "discriminant power" of each single feature separately (not e.g. feature combination), according to class separability criteria the discriminability of i[th] feature is represented by i[th] row in feature matrices M_{F1}^{p} or M_{F1}^{q} (depending on class type).

III. Define $DM(p_i, q_i)$ as a discriminate measure for the i[th] feature, which expresses what is the value of separability weight of this given feature in classification process.

Different type of $DM(p_i, q_i)$ can be considered and several was tested [6]:

a) Fisher's class separability index:

$$DM(p_i, q_i) = \frac{(mean(p_i) - mean(q_i))^2}{\mathrm{var}(p_i) + \mathrm{var}(q_i)} \tag{3}$$

where mean(·) and var(·) are computed across i[th] matrice row.

b) Relative entropy: $DM(p_i, q_i) = p_i \log \frac{p_i}{q_i}$ (4)

c) Euclidean distance: $DM(p_i, q_i) = \left\| p_i - q_i \right\|$ (5)

IV. Calculate $DM(p_i, q_i)$; $i = 1..M$ for every of M primary features.

(6)

V. Sort obtained in III $DM(p_i, q_i)$ values to create a feature rank as a results of feature competition.

VI. Choose the most discriminant *L* features to create a new feature vector X_{F2}

D. Adaptive Resonance Theory Network (ART2)

Adaptive resonance theory (ART) introduced by Carpenter and Grossberg [7] can be used for classifying binary inputs. Next structures of this family ART 2 and fuzzy ART were modified to allow to deal with analog input signals (e.g. gray scale picture). The architectures of all these networks were based on the idea of adaptive resonant feedback between two layers of nodes as developed in [7]. The ART2 network consists of three main components, termed by Carpenter and Grossberg as the input representation field (or F1 layer), the category representation field (or F2 layer) and the orienting mechanism (fig.2). The nodes of the output layer $\bar{y}$ are intended to represent the classes into which input patterns are organized. Each output node represents a class of patterns by storing a template pattern as the weights matrix: $W = \left(\bar{w}_1, \bar{w}_2, .., \bar{w}_n\right)^T$ on its connections into the F1 layer. During an ART2 trial, the input pattern is matched against each of the stored templates, resulting in activations at the output layer which represent the extent of match. At this stage of processing the output layer activations are established as the weighted sum of a layer $\bar{x}$ within the representation field F1 (7):

$$\bar{y} = \bar{x} * W^T \tag{7}$$

The particular activation value y_j is a measure of membership of the current input vector $\bar{x}$ to j^{th} class.

The element with the highest activation value - y_r designated as the initial choice of class for the input pattern. The results of this competition, usually denoted as $g(\bar{y})$ presents a value *d* ($0 < d < 1$ for a given network) for the winning node, and zero otherwise. The input pattern and the template for the initially chosen node are subjected to a further comparison in the orienting mechanism, and if the match is judged to be within the vigilance level - y_{TR}, the initial choice y_r is taken as final and the template for that winning node is updated to more strongly represent the current input. The adaptation of the weights connected with the class *r* is performed as follows (8).

$$w_{i,r}^{(k+1)} = w_{i,r}^{(k)} + \eta(x_i, w_{i,r}^{(k)}) \text{ , where}: i = 1..n. \tag{8}$$

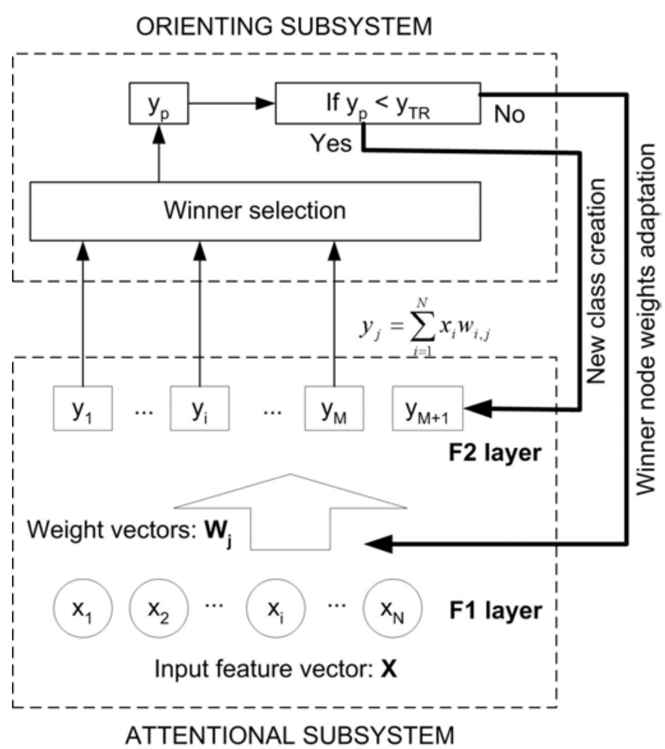

Fig. 2. Adaptive Resonance Theory Network (ART2) structure. Attentional part with input F1 and output F2 layer and orienting subsystem is presented.

If the match is judged to be outside of the vigilance level: $y_r < y_{TR}$, then the new class is added to the system.

III. RESULTS

Proposed structures were tested using the set of clinically characterized heart rate variability (HRV) signals of 62 patients, as cases with a coronary artery disease of different level. Additionally similar control group of healthy patients was analyzed. Whole database was divided into learning and verifying set. Classification task was defined as the trial of two group (healthy and pathology cases) distinguish, based on new feature subsets obtained in feature extraction and selection stages of presented procedure. In first result group a normalised discriminate measures $DM(p_i, q_i)$, computed for each feature from assumed HRV based feature group. Based on obtained feature rank, the number of most important

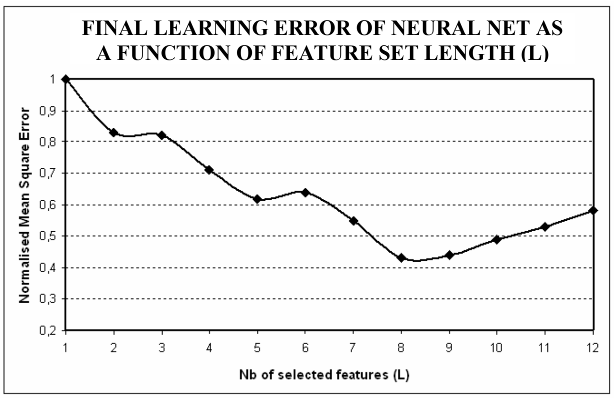

Fig.3 Neural network classifier training error as a function of number of chosen features in new feature vector.

features were taken as eight (fig.3).

Finally whole presented classifier system were verified using test set of HRV data from patient with coronary artery disease. Common used sensitivity and specificity classifier performance measures obtained for both unsupervised ART2 and supervised Multi-Layer Perceptron learning structures with and without preliminary feature extraction stage are presented in fig.4.

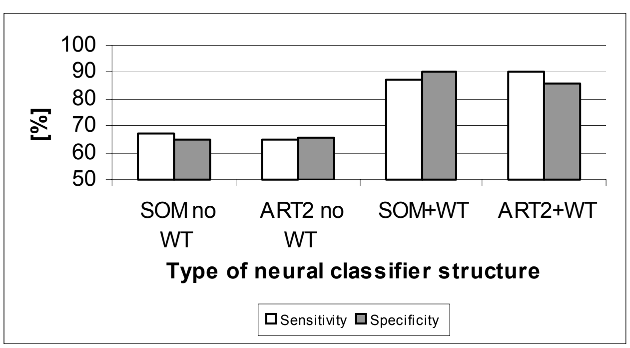

Fig.4 Neural network based classifier performance MLP and ART2 neural classifier with and without wavelet feature extraction stage.

IV. DISCUSSION

Evaluation of all extracted HRV features based on the measure of its class separability property showed, that the most discriminant features for given classification task are the parameters in vectors $\boldsymbol{F1_{C3}}$, $\boldsymbol{F1_{C4}}$ and $\boldsymbol{F1_{C5}}$ feature vectors. These feature vectors include energy, entropy and Shannon entropy E of energy distribution parameters respectively of Mallat HRV signal decomposition components. The most significant features from these vectors are assigned to $d3$ and $d4$ levels of discrete wavelet analysis (frequency subbands: $0.3125 \div 0.6250$ and $0.1563 \div 0.3125$ [Hz]). It corresponds to rather high frequency (HF) components of HRV PSD function. The comparison of supervised (MLP) and unsupervised (ART2) learning neural network used as classifier tools gave similar results but the most important is clearly seen in fig.4 positive influence of feature extraction and selection stage on classifier performance.

V. CONCLUSION

Obtained results showed, that before pattern classifier can be properly designed and effectively used, it is necessary to consider the feature extraction and data reduction problems. Feature extraction should consists in choosing those features, which are most effective for preserving the class separability.

REFERENCES

[1] M. Akay, Time-Frequency and wavelet analysis, IEEE EMB Magazine 14(2), 1995.

[2] K. Fukunaga. Introduction to statistical pattern recognition. 2nd edition. Academic Press, San Diego, CA, 1990

[3] R.O. Duda, P.E. Hart, Pattern classification and scene analysis, John Wiley & Sons, New York, 1973.

[4] P. Kostka, E. Tkacz, The comparison of the wavelet-neural systems and radial neural networks in classifier structure,16th Intern. Eurasip Conf. Biosignal2002, pp.46-48,June 2002, Brno,Czech Republic.

[5] L. Karhunen, J. Joutsensalo, Generalizations of principle component analysis, optimization problems, and neural networks, Neural Networks, vol. 8, No. 4, pp. 549-562, 1995.

[6] S. Kullback, R.A. Leibler, On information and sufficiency, Annals of Mathematics and Statistics, 22, pp. 79-86, 1951.

[7] G.A Carpenter, S. Grossberg, ART 2: Self-organisation of Stable Category Recognition Codes For Analog Input Patterns. Applied Optics, 26, P. 4919-4930.